

Design of a Transcription Tool for the Kelabit Community of Bario, Sarawak

Emmy Dahliana Hossain, Sarah Flora Samson Juan, Jane Labadin and Priscillica Agas
*Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak.
hedahliana@unimas.my*

Abstract—This paper describes the design of a transcription tool developed for the Kelabit community of Bario, Sarawak based on the community's requirements and feedbacks. Transcribing is the process of making a full written copy of spoken or dictated material. Using this transcription tool, users can listen to audio files of recordings and proceed to do the transcribing within the tool itself. Afterwards, they can save the transcription into text files for future uses. Users can control the audio file while it is playing, and do updating or editing to the transcription. This project was carried out on the motivation to assist in language preservation works as Kelabit is an under-resourced language.

Index Terms—Transcription Tool; Kelabit Language; PRISMA.

I. INTRODUCTION

Software users have expectations when using software, and when their expectations are not met, users may find it difficult to use the software, and may abandon using the software out of frustration with the difficulties encountered. The digital divide between urban and rural population could also affect the way users handle software – for those who are used to using computer technology, they may find it easier to navigate around when using software. The users who do not use much computer technology may not be able to do so. Hence the researchers of this project believe that it is important to obtain user requirements, in order to find out the users' expectations and their level of computer technology proficiency, before designing and developing the software for the target group of users.

Transcribing is the process of making a full written copy of spoken or dictated material (transcriptions). In computer science, transcribing is the process of transferring information from one recording and storing system to another. This involves using a type of software called transcribing tool to do the transcription work: the spoken or dictated material that is in audio file is listened to by the person doing the transcribing (called the transcriber), who will then type the transcription text on the tool's text editor. This will then be saved, and the transcription can be stored in the form of text files for further usage in other endeavours. The transcriptions stored can be considered as language resources that can be used in human language technologies such as machine translation, speech recognition etc. These technologies require large amounts of resources (data) in order for it to be developed and move forward. With the availability of these resources, it is hoped the growth of language technologies can be spurred on. In time, the language can be preserved, documented and ultimately – considering its threatened status – prevent its extinction.

There are 63 known indigenous languages in the Malaysian state of Sarawak on Borneo island. One of these languages is the Kelabit language, spoken by the Kelabit community, which originates from the remote mountainous region of Bario in northern Sarawak. The Kelabit language is considered an under-resourced language, which means the language do not have sufficient resources to maintain itself [1]. Examples of language resources are speech and text data, linguistic resources (including computational linguistic resources such as corpora, pronunciation dictionaries, vocabulary lists, grammars, taggers, morphological analyzers, parsers, speech recognizers, annotation standards and tools etc.), and literature as well as language technologies (automatic speech recognition, machine translation, etc.) [2]. In addition, there are not many linguistic experts in this language, as well as the language lacks presence on the World Wide Web. All these contribute to the language being under-resourced, and may lead it to extinction if there are no efforts to preserve it in digital formats, and if no language technologies are implemented on time for the language. As of August 2016, Ethnologue [3] classifies the Kelabit language as "Level 6b – Threatened" on the Expanded Graded Intergenerational Disruption Scale (EGIDS). This denotes that while the language is spoken within all generations, it is facing declining usage among its speakers, especially among the youth. Some members of the older generation in Bario have stated in the interviews that the Kelabit language spoken by their younger generation is either "diluted", "mixed" or "not the pure Kelabit language anymore", which comes as a result of assimilation with other languages such as Malay and English.

This paper describes the efforts undertaken to design a transcription tool software for the Kelabit community in Bario, Sarawak, Malaysia. The rest of the paper is divided into the following sections: related works, method, discussion, acknowledgements and references.

II. RELATED WORKS

Transcribe is a simple Chrome web app that allows audio uploading and the transcribing process does not need the switching between the audio player and the text editing document. Transcribe is preferable in Journalism.co.uk such that they opt to record Skype and phone interviews to focus on the conversation and refer back in other time [2]. This tool is a time-saving tool as user can transcribe quotes or sections from an audio interview and toggle between the audio player and the text-editing document. This web app offers an audio player which is integrated with a text editor on the screen interface. User can upload an mp3, mp4, m4a, amr, wma, aac,